Syntactic complexity variation between Dutch learners of English in CLIL education as opposed to regular education

BA Thesis: Linguistics

Marleen Roncken - s4494237

English Language and Culture - Radboud University

Abstract

This study investigates the variation in syntactic complexity in written essays by Content and Language Integrated Learning (CLIL) students as opposed to non-CLIL students in The Netherlands. All participants are Dutch learners of English who are in their final year of a secondary school in The Netherlands. The analysis is based on 40 argumentative essays from students in both CLIL and non-CLIL programmes. Based on previous research, eight measures were chosen to assess syntactic complexity and the essays were analysed using Lu's Second Language Complexity Analyzer. Results reveal significant differences in mean length of clause, coordinate phrases per T-unit and coordinate phrases per clause, indicating that CLIL students generally use longer clauses, and more coordinate phrases in argumentative essays. The other syntactic complexity measures did not show any significant results. This study concludes with possible implications and limitations, with points of focus for future research.

Keywords: CLIL education, CLIL programme, L2 writing, EFL, syntactic complexity, secondary education

Table of contents

Introduction

Content and Language Integrated Learning (CLIL) is an educational programme that has been increasing in popularity in the past couple of decades. While the exact properties of CLIL vary between countries and even between educational institutes, the general goal remains the same, which is the acquisition of a foreign language through teaching other subjects in this language. In The Netherlands, there are over 130 schools that have CLIL programmes ("Alle tto-scholen in Nederland", n.d.), with English usually being the foreign language that is being integrated. Many of the subjects taught in those programmes are in English, so students would not only learn the subject, but they would also become more proficient in the foreign language.

CLIL-programmes have generally become more popular because of the assumption that bilingual education increases the students' second language proficiency in that second language. However, there is limited research that clearly demonstrates that this is actually the case. Therefore, more research is needed to investigate the effectivity of CLIL programmes by examining the L2 proficiency of CLIL students relative to non-CLIL students.

There are various ways to test proficiency of a second language. Many studies examining L2 proficiency and development use the CAF framework as a basis, which categorises proficiency into three main traits: complexity, accuracy, and fluency. This study focuses on the syntactic complexity of written texts produced by CLIL students as opposed to texts written by students attending regular education. In order to obtain relevant results, the syntactic complexity of essays from CLIL and non-CLIL students will be analysed and compared. The findings of this study findings could provide more clarity on the effectivity of L2 acquisition through CLIL education in The Netherlands.

CLIL and bilingual education

Even before the term Content and Language Integrated Learning (CLIL) was coined by David Marsh in 1994 (Biçaku, 2011), researchers had attempted to examine the effects of foreign language instructional settings in second language proficiency. Hamayan and Tucker (1980) analysed the effect of bilingual language use in the classroom. The paper considered two groups located in Montreal: one group consisted of bilingual classes taught by a bilingual teacher (English L1/French L2) while the other group consisted of monolingual French students taught by a French Canadian

teacher. A recorder was used to observe utterances by both the teacher and the students. The students within immersed bilingual settings tended to replicate correct utterances more than the non-bilingual setting students. The results showed that the teachers in bilingual settings tend to use teaching strategies that require more input from the student, such as asking open questions, which may be one of the reasons why the bilingual students performed better in general (pp. 466-67).

However, after its introduction, CLIL quickly became a topic of interest due to its rising popularity at educational institutes and because of concerns related to effectivity of acquiring a foreign language through the programme. In order to examine those concerns, multiple studies have tried to assess L2 proficiency of CLIL students relative to non-CLIL students. De Zarobe & Catalán (2009) analysed the receptive vocabulary proficiency of EFL learners in CLIL and non-CLIL programmes. 130 participants (Spanish learners of English) used for this study were tested used the 1000-word receptive test and the Vocabulary Levels Test. The CLIL students performed significantly better than the non-CLIL students in both tests, suggesting that the receptive vocabulary skills of CLIL students were better. In a later study, de Zarobe (2010) aimed to analyse the language proficiency differences between CLIL and non-CLIL students, but this time with a focus on written production. It was a quasi-longitudinal study, which compared a third-year and a fourth-year secondary school group, along with a pre-university group. All students were asked to write a letter with specific instructions which were then assessed based on five categories (content, organisation, vocabulary, language usage, and mechanics) and using four rating levels. The results show that CLIL students in all education levels scored significantly higher in all categories. Most studies that explore the effectiveness of CLIL in relation to L2 proficiency generally found significant differences that suggest that students' L2 proficiency benefit from CLIL programmes in multiple language areas such as vocabulary and writing.

Although results of studies comparing CLIL and non-CLIL education are often interpreted in a way that CLIL leads to better foreign language development, this connection may not always be true. Möller (2017) examined individual factors and differences that affect the L2 success in CLIL and non-CLIL programmes. She analysed the use of passive constructions amongst learners that voluntarily choose for or opt out of a CLIL programme, along with learners that did not have a CLIL programme available to them. The findings confirmed a certain inherent selectivity

when it comes to learners opting for a CLIL programme. CLIL learners showed significantly higher language skills than non-CLIL students in general, implying that CLIL learners already have a head start in L2 development. CLIL students also engaged in more out-of-class activities related to the English language. It seems that it is possible that students of CLIL programmes may be proficient because of reasons outside of the programme itself. Another study that explored individual factors influencing L2 proficiency in CLIL and non-CLIL programmes came to a similar conclusion (Verspoor et al., 2015). Two groups of Dutch high school students were tested three times within one year, while also analysing interplay from factors such as initial proficiency, scholastic aptitude and motivational factors. The CLIL group outperformed the non-CLIL students, but it could not be concluded whether L2 proficiency was linked to the higher amount of L2 input in CLIL teaching. Factors such as initial proficiency, scholastic aptitude, motivation, and out of school contact were considered to be factors that influenced L2 success. The findings suggest that students who opt for a CLIL programme are generally more motivated and have more linguistic aptitude, which are factors that could explain why CLIL students are generally more proficient in their L2 relative to non-CLIL students. There are not many studies that investigate these factors more elaborately yet, so we cannot assume that this is the case in every context. However, the possibility of other factors at play should be recognised when researching L2 proficiency in CLIL and non-CLIL programmes.

In conclusion, there is some evidence that CLIL programmes may positively stimulate L2 language proficiency. However, as many of these studies mention, more research towards the effectivity of CLIL education on second language development is still desired.

CLIL in the Netherlands

CLIL schools have also been on the rise in The Netherlands. English is taught alongside Dutch in a more immersed manner than in regular education schools, causing Dutch L1 speakers to become more familiar with English than at regular secondary schools in the Netherlands. Admiraal et al. (2004) investigated CLIL secondary education in The Netherlands and how it affects EFL proficiency. All six levels within secondary education were analysed, with a total of 584 students participating. In order to measure language proficiency, the English as a Foreign

Language Vocabulary Test (EFL Vocabulary Test) was used, along with a reading comprehension test with multiple-choice questions and an oral proficiency test. Results showed that students would significantly improve on their reading ability throughout the levels of education, but that their vocabulary growth remained similar to students in regular education. However, Admiraal et al. stressed that this research was done at a 'pioneer' school when a set curriculum for their bilingual programme had not been established yet. Since bilingual education is a relatively new concept in The Netherlands, Admiraal et al. encouraged other researchers to do more research on the subject. Van Daalen (2016) also researched the effectiveness of CLIL education in The Netherlands. Twenty students (ten from CLIL education, ten from regular education) were picked and asked to do a picture-description exercise, which was evaluated by several proficient English speakers. The results show that CLIL students seem to produce more fluent sentences, forming longer sentences more spontaneously with better pronunciation than students attending regular education (p. 23). Van Daalen theorised that this is due to prolonged English (L2) exposure in the classroom and many interaction moments during class.

Although there is some research investigating differences between, CLIL and non-CLIL programmes in The Netherlands, there are not many studies that explore the differences in EFL proficiency yet. More insight on the effectivity of CLIL programmes in regards to EFL success will be helpful in determining whether the CLIL programmes in the Netherlands are effective in their current state.

CAF framework

Before proficiency in CLIL and non-CLIL programmes can be analysed, the question that needs to be asked is how a learner's second language proficiency can be measured. Much research in the past tackled this topic, and the consensus is that second language proficiency is multi-dimensional in nature. Second language acquisition seems to consist of three major components, complexity, accuracy, and fluency, which have been used by researchers since the 1970s (Housen et al., 2012). However, these proficiency dimensions have always been discussed with little relation to one another. In the 1990s, complexity, accuracy, and fluency were considered together as a way of measuring L2 proficiency (Skehan, 1996). This is how the CAF (complexity, accuracy, fluency) framework was formed, which serves as a basis for measuring L2 proficiency in SLA studies and instructional institutes. To

this day, the three dimensions appear, simultaneously or individually, in L2 proficiency research.

Accuracy has generally been a stable term throughout L2 research and refers to the rate of 'correctness' of an utterance produced by an L2 learner compared to how a native speaker would utter it (Housen et al., 2012). How much a learner deviates from the norm (native speaker) decides the level of accuracy. Most of the time, the researchers that examine accuracy of L2 learners either try to identify the number of accuracy errors that have been produced or they attempt to measure how much of the utterance is free of errors (Lambert & Kormos, 2014, p. 609). Although measuring correctness is a seemingly straightforward way to determine L2 proficiency, there are some issues regarding reliability and subjectivity that quickly arise. The core of the problem is that determining when something is 'right' or 'wrong' is often a very difficult thing to assess because there is not always a definitive answer. The degree of an error is usually not taken into account either.

While fluency is a more complex performance area to describe, it generally revolves around how 'smoothly' the L2 learner can produce the language based on sentence constructions, pronunciation, vocabulary, and speed of production (Mitchell et al., 2017, pp. 28-28). Fluency can be measured in both speech and writing tasks, but speech is often preferred because variables such as response time, hesitance, and self-correction can be measured with ease. Because fluency has many variables, researchers have tried to break down the performance area into multiple categories. Examples of these are the three categories defined by Segalowitz (2010), dividing fluency into cognitive, utterance, and perceived fluency (as cited in Tavakoli et al., p 449).

Complexity has various definitions within L2 research, but in SLA research, it is generally split up into two subcategories: cognitive complexity and linguistic complexity (Housen & Kuiken, 2009, p.462). While cognitive complexity has a focus on factors dependent on the L2 learner (e.g. L1 background, motivation), linguistic complexity focuses more on what the learner produces in terms of "size, elaborateness, richness, and diversity" (p. 464). The linguistic complexity subcategory can be broken down even further, but the way this is done usually varies depending on the study. Many studies make the distinction between lexical complexity, which refers to the size and depth of the L2 vocabulary of a learner, and syntactic complexity, which refers to structural complexity on a sentential, clausal and phrasal

level. Lexical complexity is normally measured by investigating the lexical density (content words) of an utterance. Syntactic complexity can be measured by examining the length of units (sentences, clauses, and T-units) or by measuring ratios of certain structures (e.g. ratio of clauses per sentence, ratio of coordinate phrases per clause).

While these performance dimensions are usually studied separately from each other, they are often connected in some way. Skehan (2009) investigated the CAF framework in relation to task-based instruction and noted that most tasks overlap in terms of performance dimensions. For example, interactive tasks will test and improve a learner's accuracy and complexity the most (p. 511). However, since it is difficult for researchers to cover multiple performance areas within one case study, researchers try to stick to investigating one performance area when measuring L2 proficiency.

The current study focuses on investigating student L2 performance by looking at syntactic complexity. Most studies that investigated CLIL and non-CLIL programmes measured fluency or accuracy, but (syntactic) complexity has not often been considered as a variable. While L2 proficiency is a challenge to measure accurately and reliably, syntactic complexity is generally considered to be a good indicator (Bulté & Housen, p. 148). One of the reasons is that the syntactic complexity measures are straightforward and used very similarly in nearly all syntactic complexity studies related to L2 proficiency. The next section will elaborate on the definitions of these measures and how they are used.

Syntactic complexity measures and L2 writing

Syntactic complexity has been widely discussed by SLA researchers attempting to pinpoint syntactic language proficiency, especially because the CAF framework suggests that language complexity is one of the major factors that determines L2 proficiency. Norris and Ortega (2009) aimed to analyse how syntactic complexity should be implemented in the CAF framework. Several studies that investigated syntactic complexity in SLA were analysed in order for Norris and Ortega to come up with suggestions that would provide a better understanding to researchers regarding what exactly is being measured or researched. When measuring syntactic complexity, research should include general complexity, complexity by subordination, complexity by coordination and complexity by phrasal or clausal elaboration. Additionally, Norris and Ortega hoped to see more syntactic complexity

measures specifically directed towards second language acquisition instead of language acquisition in general. Furthermore, they wanted to see a more unified method of reporting and calculating complexity measures such as measuring subordination and coordination, while also adding in and interpreting factors that could influence the research. This study implies that with these suggestions, syntactic complexity in foreign language learning can indeed be measured in relation to the CAF framework.

Since syntax has many elements that are often too broad to be studied altogether, there are many studies that aim to answer the question if and how syntactic proficiency can be accurately measured and how it can be linked to L2 proficiency. Many researchers have attempted to investigate syntactic complexity on a smaller scale, only looking at specific normalised elements that may help measure syntactic complexity. Table 1 presents an overview of syntactic complexity measures that have been used by various researchers over the years. Ortega (2003) sought to combine past studies that made use of syntactic complexity measures and analyse them on how they assess differences in various types of L2 writing, looking specifically at the differences between foreign and second language learning. Twenty-one studies were investigated on the use of syntactic complexity measures in foreign language writing. Mean length of sentence (MLS), mean length of T-unit (MLT), mean length of clause (MLC) and clauses per T-unit (C/T) were found to be helpful in determining the syntactic complexity difference between L2 groups. However, she also mentioned that there were not enough studies yet that made use of specific complexity measures. For example, she stated that there were not enough studies and large enough groups that were tested on her other two tested syntactic complexity measures, dependent clauses per clause (DC/C) and T-units per sentence (T/S). It is also quite difficult to compare studies because the studies that were analysed in this paper all varied in L1, level of L2, age group, type of data, and which syntactic complexity measures were used. However, Ortega's research set a basis for the quest to find a set of measures that could be used to reliably determine syntactic complexity in L2 utterances.

Table 1

Overview of syntactic complexity measures used in research mentioned in this paper.

Abbreviation	Syntactic complexity measure
MLS	Mean length of sentence
MLT	Mean length of T-unit
MLC	Mean length of clause
C/S	Clauses per sentence
C/T	Clauses per T-unit
T/U	T-units per sentence
VP/T	Verb phrases per T-unit
DC/C	Dependent clauses per clause
DC/T	Dependent clauses per T-unit
CT/T	Complex T-unit per T-unit
CP/T	Coordinate phrases per T-unit
CP/C	Coordinate phrases per clause
CN/T	Complex nominals per T-unit
CN/C	Complex nominals per clause

Biber et al. (2011) questioned whether measuring T-units, a syntactic measure examined in Ortega (2003), is actually helpful in determining syntactic complexity in L2 writing development. They claim that, while a lot of research has been done on syntactic complexity using T-units as measures, there is little empirical evidence to support whether such measures are actually accurate in determining language development. For this study, Biber et al. (2011) specifically examined measures of T-units. Using syntactic tagging software, they analysed two large corpora: one consisting of academic writing from various fields (predominantly research articles) and the other a collection of face-to-face conversations, with 1.9 million and 4.1 million words, respectively. They concluded that MLT may not be indicative of actual language proficiency. Yang (2013) expressed his concern with their claim that T-units are not an accurate way of measuring writing development and stressed that Biber et al. only looked at corpus data from already proficient speakers. He pointed out that studies that examined L2 proficiency such as Ortega (2003) successfully provided evidence for the link between subordination (which can be measured in T-units and clauses) and writing proficiency. Yang mentioned that Biber et al. (2011) actually

reported a high number of non-finite subordination in their written data, adding to the importance of investigating T-units to measure writing development. In response, Biber et al. (2013) emphasised that it is hard to compare conversational corpus and formal writing with T-units alone. They explain that T-unit measures used in previous research were defined to only include finite forms, meaning that non-finite forms were excluded. It seems that T-units may not be an accurate form of syntactic complexity measurement by itself, but they may still be a helpful tool in measuring writing development in combination with other syntactic complexity measures that focus more on non-clausal syntactic properties, such as sentential length and phrasal units.

Unsurprisingly, the studies in this field that followed attempted to include more syntactic complexity categories and tested them for their reliability to measure syntactic complexity in L2 writing. Lu (2010) argued that, in order to reliably measure syntactic complexity and potentially link it to L2 proficiency in writing, sentential, clausal, and phrasal measures should be considered. He was looking to develop an automated system that would be able to accurately measure syntactic complexity in studies related to L2 development. According to him, many previous studies only looked at a small part of syntactic complexity with small amounts of data, leading to possibly unreliable results (pp. 475-476). Lu specifically mentioned Ortega's (2003) research, which investigated previous research that generally analysed only one to three syntactic complexity measures. For this analysis, fourteen measures of syntactic complexity were collected that have been widely used in previous research and were divided into five categories: clausal and sentential length (MLS, MLC, and MLT), sentence complexity ratio (C/S), amount of subordination (C/T, CT/T, DC/C, and DC/T), ratios of coordination (CP/C, CP/T, and T/S), and a category that summarises the relationship between particular syntactic units (CN/C, CN/T, and VP/T) (p. 478). See Table 1 for the definitions of these syntactic complexity measures. Note that these categories used syntactic complexity measures from previous research and also include complex phrases, which Biber et al. (2011) commented there was a general lack of in syntactic complexity research.

In a follow-up study, Lu (2011) aimed to provide evidence on whether his fourteen proposed syntactic complexity measures could indicate EFL development and how these are affected by writing genre. A random sample of 3678 essays from the Written English Corpus of Chinese Learners was collected, after which they were

analysed by Lu's L2 Syntactic Complexity Analyzer, which is an automated parsing software that can annotate syntactic structures. The results were compared with the work of experienced annotators in order to ensure accuracy. Both argumentative and narrative essays were included in this study so that their syntactic complexity could be compared as well. The results indicated that seven out of the fourteen syntactic measures showed a positive linear progression throughout most proficiency levels. The structures that showed the more drastic changes were complex nominals per clause (CN/C) and mean length of clause (MLC), followed by complex nominals per T-unit (CN/T), mean length of sentence (MLS), and mean length of T-unit (MLT), coordinate phrase per clause (CP/C) and coordinate phrase per T-unit (CP/T). (p.56). Three syntactic structures, however, showed a decrease in usage at higher levels: clauses per sentence (C/S), dependent clauses per clause (DC/C), and dependent clauses per T-unit (DC/T). These results are in accordance with Ortega (2003) and Biber et al.'s (2011) research in which they claim that higher proficiency is also linked with complex phrases rather than the usage of dependent clauses and T-units alone. Syntactic complexity in timed argumentative essays developed mostly linearly with school level.

Casal and Lee (2019) also used Lu's L2 Syntactic Complexity Analyzer to explore the relationship between syntactic complexity and writing quality in assessed L2 writing. They analysed 280 research papers that had been divided into three graded categories, namely high, mid, and low and used five holistic measures of syntactic complexity, MLT, C/T, TU/S, MLC and CN/C, divided into global, clausal and phrasal syntactic complexity. They found that essays that were categorised as having a high writing quality contained higher global (MLT) and phrasal (MLC and CN/C) complexity.

Bi and Jiang (2020) performed a similar study that aimed to research the connection between syntactic complexity and writing quality, but they were more interested in young adolescent learners (ages 10 to 19) of English. For this study, 410 narrative essays of Chinese-speaking EFL learners were used, and the essays were analysed for seven complexity measures (MLC, MLT, MLS, C/T, T/S, CP/C, CN/C). These measures were then used to test the reliability of variance in writing scores. Bi and Jiang concluded that the measures mean length of sentence (MLS), complex nominals per clause (CN/C) and clauses per T-unit accounted for 36,6% of the variance in writing scores and were shown to be reliable (p. 8).

Table 2

General overview of the results relating to syntactic complexity measures mentioned in this section.

Research	*MLS*	*MLT*	*MLC*	*C/S*	*C/T*	*DC/C*	*DC/T*	*T/S*	*CP/T*	*CP/C*	*CN/T*	*CN/C*
*Ortega (2003)**	+	+	+		+	+/-		+/-				
Biber et al. (2011)		-				-	-	-				
Lu (2011)	+	+	+	-	-	-	-	-	+	+	+	+
Casal & Lee (2019)		+	+									+
Bi & Jiang (2020)	+	-	-		+			-	-			+

Note: + = positive evidence in research; - = negative evidence in research, +/- = inconclusive data, empty slots refer to syntactic complexity measures that have not been included.

Table 2 shows a general overview of the research mentioned in this paper relating to testing the reliability of syntactic complexity measures of L2 written texts. Note that some of Lu's proposed syntactic complexity measures, VP/T and CT/T, have been left out as they have not been taken into account in other research on syntactic complexity. This overview indicates that there is no conclusive list of syntactic measures that can consistently and accurately determine L2 writing quality yet. There are some syntactic complexity measures that have generally produced positive correlations. The clausal and sentential length measures produced reliable results in Ortega (2003), Lu (2011), Casal and Lee (2019), and Bi and Jiang (2020). There is also some evidence that MLC and MLT may not be reliable, however, so these measures should be carefully interpreted. Complex nominals per clause (CN/C) seems to be generally reliable when determining writing quality in Lu (2011), Casal and Lee (2019), and Bi and Jiang (2020), along with clauses per T-unit (C/T) in Ortega (2003) and Bi and Jiang (2020). While coordinate phrases per T-unit (CP/T), coordinate phrases per clause (CP/C), and complex nominals per T-unit (CN/T) were considered to be reliable by Lu (2011), there is generally not much evidence towards using them in L2 syntactic complexity research. CN/T has not shown evidence of reliability in Bi & Jiang and the CP/C and CN/T have not been used in other research, so further research using these measures would be beneficial. The syntactic

complexity measures clauses per sentence (C/S), dependent clauses per clause (DC/C), dependent clauses per T-unit (DC/T), and T-units per sentence (T/S) have generally not shown to be reliable when measuring L2 writing quality.

Present study

The present study examined the differences in L2 language proficiency in bilingual education as opposed to regular education in the Netherlands. More specifically, it attempted aimed to analyse the differences in the syntactic complexity of EFL writing. In accordance with previous research, this paper used eight syntactic complexity measures: mean length of sentence (MLS), mean length of T-unit (MLT), mean length of clause (MLC), clauses per T-unit (C/T), coordinate phrases per T-unit (CP/T), coordinate phrases per clause (CP/C), complex nominals per T-unit (CN/T) and complex nominals per clause (CN/C) (cf. Table 3). Not all fourteen syntactic complexity measures used in previous research have been shown to be reliable to measure L2 writing quality and proficiency. MLS, C/T, and CN/C have been shown to be consistently reliable measures for determining L2 writing quality. MLT, MLC, CP/T, CP/C, and CN/T have yielded mostly positive results, but there is not enough evidence yet to consider these syntactic complexity measures as consistently reliable, which needs to be born in mind when interpreting the results.

Table 3

Overview of the syntactic complexity measures that are used for this study.

Abbreviation	**Syntactic complexity measure**
MLS	Mean length of sentence
MLT	Mean length of T-unit
MLC	Mean length of clause
C/T	Clauses per T-unit
CP/T	Coordinate phrases per T-unit
CP/C	Coordinate phrases per clause
CN/T	Complex nominals per T-unit
CN/C	Complex nominals per clause

Since many previous studies show that higher syntactic complexity correlates with L2 writing quality, this study aims to assess whether CLIL students will also

show higher syntactic complexity in their writing as opposed to non-CLIL students. The research question addressed by this study is:

> How does CLIL education as opposed to regular education affect syntactic complexity in the writing of Dutch L2 learners of English?

CLIL students generally scored higher than non-CLIL students in terms of L2 proficiency in previous studies, I also expect the CLIL group to write more syntactically complex L2 texts.

Participants and materials

All participants recruited for this study are students from the Kandinsky College in Nijmegen. The participants are divided into two groups: that that follows regular secondary education and one that follows CLIL secondary education, also known in The Netherlands as TTO (*tweetalig onderwijs*). The school offers both types of education simultaneously, so when students enroll they may choose whether to follow 'regular' classes or bilingual classes. In the bilingual classes, similar to all CLIL education in The Netherlands, more than 50% of subjects are taught in English to create immersion and familiarity with the language in the first three years of their secondary education. From the fourth year onwards, more classes are taught in Dutch again to prepare students for the central exams, which take place in the final year and are the same across all secondary schools in The Netherlands, regardless of the type of education.

All students that participated were in their final year of secondary school and were all at the same educational level (6 VWO). Following the Common European Framework of Reference for Languages (CEFR), the exit level of English proficiency of students in this education level is B2. Students from both regular education and bilingual education were randomly picked to form two equal groups of twenty (40 in total). The data is collected from hand-written argumentative essays that were written independently from this research. The students were asked to write a 500 to 700 word-essay within a 100-minute time limit on an assigned topic. The students were allowed to choose from a handful of political, economic, or personal statements that they had to use as a basis for their essay. Other than that, the students were allowed to write whatever they wanted, so the texts contain natural utterances in a relatively formal register. Students did not know about the topics beforehand, but they were allowed to use a dictionary during the exam.

Table 4 shows an overview of the essays that had been collected. The twenty essays had been randomly picked from a larger pool of students. Besides the total word count of both datasets, the table also shows the mean length of each essay per group, along with the standard deviation.

Table 4

Summary of the dataset used for this study.

Education type	Number of essays	Total word count	Mean length per essay	Standard deviation (SD)
Regular	20	11077	553,85	102,03
CLIL	20	12437	621,85	98,50

It is important to note that at the time of writing this thesis and collecting the data, the COVID-19 pandemic was active in The Netherlands. As a result of this, many educational institutions, secondary schools included, were heavily impacted. All schools, including the Kandinsky College, were closed for over a month and were forced to cancel the central exams that, under normal circumstances, take place nationally in spring. The school used an essay assignment that was treated as a subsidiary method of examining the students, which are also the essays that were used for this study.

Analytical tools

The syntactical measures were analysed and calculated by Lu's L2 Syntactic Complexity Analyzer (L2SCA). Lu (2010) developed this software to automate measurements of syntactic complexity. There are already existing syntactic complexity measuring programs, but they generally require the researcher to also manually tag each sentence while this software annotates texts automatically. The L2SCA that uses the Stanford Parser, which tags syntactic structures within a written text, from which Lu's software extracts syntactic complexity values. To test the reliability of the L2 Syntactic Complexity Analyzer, 40 essays were randomly picked from the Written English Corpus of Chinese Learners, which were then run through the software and also checked by experienced annotators. The results showed that the Stanford parser determines the syntactic structures with high accuracy with F-scores ranging from 0,834 to 1,000 (p. 491), which means that this software may be helpful with determining syntactic complexity in L2 writing. Since Lu used the Stanford Parser software to determine syntactic complexity in L2 data, he was also interested in researching to what extent the fourteen measures he used can indicate L2 language development (2011). He argued that a lot of researchers have been hesitant to research syntactic complexity due to the lack of automated tools, and thus decide to work with

smaller amounts of data or limited syntactic complexity measures (p.37). His study aimed to analyse whether an automated system using the Stanford parser can accurately measure specific syntactic structures and if the results can be linked to proficiency. For the results, Lu once again used his software (L2SCA) and experienced annotators to obtain the results, and similar to his 2010 study, the software was very accurate compared to the human annotators.

Lu's L2SCA is not the only syntactic complexity annotator available, however, but comparisons show that the L2SCA is most effective tool to examine syntactic complexity in L2 writing. Lu (2017) examined and compared three systems that measure syntactic complexity in L2 data. He summarised the properties of the Biber Tagger from Biber, Leech, Conrad & Finegan (1999), Coh-Metrix by McNamara, Graesser, McCarthy & Cai (2014) and Lu's L2 Syntactic Complexity Analyzer (2010). While the Biber Tagger and Coh-Metrix are also able to annotate syntactic complexity in L2 writing, these two analysers were not designed to be used for this purpose specifically, so their calculated datasets are generally not as complete as the one generated from the L2SCA.

Several case studies have provided evidence that Lu's L2 Syntactic Complexity Analyzer is accurate. Casal and Lee (2019) and Bi and Jiang (2020) both used the L2SCA in order to measure syntactic complexity in L2 writing in order to measure writing quality. These studies provided evidence for the link between syntactic complexity and L2 writing quality.

However, not all syntactic complexity measures that can be measured in Lu's L2SCA have shown evidence of being reliable when determining syntactic complexity in L2 writing. Casal and Lee (2019) and Bi and Jiang (2020) made a selection of the fourteen syntactic complexity measures that they deemed to be relevant and reliable for their research. Similar to these studies, this paper used a selection of complexity measures that have been shown to be effective in measuring syntactic complexity in L2 writing. Table 5 shows an overview of the chosen measures.

Procedure

All essays were anonymised and digitised by hand after being categorised by education type. Mean length and standard deviation of all essays were calculated for both groups. After that, each essay was analysed separately in the L2SCA software, and the eight measures compiled in Table 5 were used.

Table 5

Overview of the syntactic complexity measures used for this study.

Abbreviation	**Syntactic complexity measure**
MLS	Mean length of sentence
MLT	Mean length of T-unit
MLC	Mean length of clause
C/T	Clauses per T-unit
CP/T	Coordinate phrases per T-unit
CP/C	Coordinate phrases per clause
CN/T	Complex nominals per T-unit
CN/C	Complex nominals per clause

These measures were also divided into three categories, sentential, clausal, and phrasal, to get a better overview of the syntactic complexity on multiple levels. The The L2 Syntactic Complexity Analyzer calculated specific complexity measures for each essay separately. All measures for both CLIL and non-CLIL texts were compared using an independent t-test for statistical significance. For this study, we assumed a standard p-value of 0.05 ($p < 0.05$). Any statistically significant results were marked for clarity.

Results

Since this study aims to determine whether there is a difference in syntactic complexity between the written essays of students in CLIL education as opposed to regular education, the results need to indicate if there is a difference in measured syntactic complexity categories and whether those differences are significant. Table 6 shows mean syntactic complexity values of all essay entries per group, along with standard deviation and significance based on an independent t-test. This section will report on eight syntactic complexity measures, divided into three categories: sentential complexity (mean length of sentence (MLS), mean length of T-unit (MLT) and mean length of clause (MLC)), clausal complexity (clauses per T-unit (C/T)), and phrasal complexity (coordinate phrases per clause (CP/C), coordinate phrases per T-unit (CP/T), complex nominals per clause (CN/C), complex nominals per T-unit (CN/T)).

Table 6

Mean values, standard deviations and significance of syntactic complexity measures of both groups.

SC msr.	***Education type***	***n***	***Mean***	***SD***	***t-test****
MLS	Regular	20	17,15234	2,959209	0,069662
	CLIL	20	18,60641	1,796676	
MLT	Regular	20	15,43215	2,444454	0,177283
	CLIL	20	16,37089	1,822776	
MLC	Regular	20	9,39178	1,141197	0,017015*
	CLIL	20	10,40208	1,400722	
C/T	Regular	20	1,65253	0,237757	0,376137
	CLIL	20	1,589785	0,203953	
CP/T	Regular	20	0,30489	0,121501	0,037509*
	CLIL	20	0,402105	0,160309	
CP/C	Regular	20	0,1866	0,081596	0,026159*
	CLIL	20	0,25406	0,101362	
CN/T	Regular	20	1,89065	0,510647	0,940768
	CLIL	20	1,901045	0,353381	
CN/C	Regular	20	1,15247	0,292962	0,500089
	CLIL	20	1,21391	0,277536	

Note: SC msr. = syntactic complexity measure; t-test = result of t-test calculation between the two groups, with * showing whether the difference is significant. Significance is based on the value of $p < 0,05$.

As can be seen in Table 7, mean length of sentence (MLS), mean length of T-unit (MLT), and mean length of clause (MLC) are all consistently longer for the CLIL group. Only the difference in mean length of clause was significant (0,017015), however.

Table 7

Mean values, standard deviations, and significance of sentential complexity measures of both groups.

SC msr.	***Education type***	***n***	***Mean***	***SD***	***t-test****
MLS	Regular	20	17,15234	2,959209	0,069662
	CLIL	20	18,60641	1,796676	
MLT	Regular	20	15,43215	2,444454	0,177283
	CLIL	20	16,37089	1,822776	
MLC	Regular	20	9,39178	1,141197	0,017015*
	CLIL	20	10,40208	1,400722	

Table 8 shows the results regarding clausal complexity in the essay texts. When it comes to clauses per T-unit, the non-CLIL students outperform the CLIL students, although that difference is not significant.

Table 8

Mean values, standard deviations, and significance of clausal complexity measures of both groups.

SC msr.	***Education type***	***n***	***Mean***	***SD***	***t-test****
C/T	Regular	20	1,65253	0,237757	0,376137
	CLIL	20	1,589785	0,203953	

The results regarding clausal complexity can be found in Table 9. Calculation of the mean values of coordination (CP/C and CP/T) shows that students of CLIL education consistently produce more coordinate phrases per T-unit and clauses. The differences between the groups in CP/T and CP/C are significant (0,037509 and 0,026159, respectively). When considering the use of complex nominals (CN/T and CN/C), the

essays from CLIL students show higher ratios of complex nominals per clause and T-units. Again, none of these syntactic measures show significant differences.

Table 9

Mean values, standard deviations, and significance of phrasal complexity measures of both groups.

SC msr.	***Education type***	***n***	***Mean***	***SD***	***t-test****
CP/T	Regular	20	0,30489	0,121501	0,037509*
	CLIL	20	0,402105	0,160309	
CP/C	Regular	20	0,1866	0,081596	0,026159*
	CLIL	20	0,25406	0,101362	
CN/T	Regular	20	1,89065	0,510647	0,940768
	CLIL	20	1,901045	0,353381	
CN/C	Regular	20	1,15247	0,292962	0,500089
	CLIL	20	1,21391	0,277536	

In summary of these findings, CLIL students generally produced longer sentences and clauses, while also using more coordinate phrases and specific phrasal units such as complex nominals. However, it is only the mean length of clauses (MLC) and usage of coordinate phrases per clause and T-unit (CP/C, CP/T) in the written texts of CLIL students that were significantly higher than those of non-CLIL students.

Discussion

This study investigated the differences in syntactic complexity in EFL writing between CLIL and non-CLIL students. The essays written by CLIL students produced higher syntactic complexity in seven out of eight syntactic complexity measures, with the exception of clauses per T-unit (C/T). Out of those seven measures, only three measures differed significantly between the two groups. The three measures that had significant variation between groups were mean length of clause (MLC), coordinate phrases per T-unit (CP/C), and coordinate phrases per clause (CP/C). The students in the CLIL programme generally seemed to produce texts that were syntactically more complex on a sentential and phrasal level compared to the non-CLIL group.

These results are in line with previous research on L2 proficiency of students in CLIL and non-CLIL programmes. Many studies that examined the L2 proficiency variations between CLIL and non-CLIL students generally found significant results that suggested that CLIL students are more proficient in their second language (Admiraal et al., 2004; de Zarobe & Catalán, 2009; van Daalen, 2016). Studies that more specifically investigated differences in L2 writing also found that the students participating in CLIL programmes are more proficient in terms of structure and vocabulary usage (de Zarobe, 2010). Although this study examined a different field within L2 writing, some of the findings still imply that CLIL students can produce more syntactically complex writing compared to non-CLIL students.

Interestingly, some of the syntactic complexity measures that needed to be interpreted carefully due to inconclusive data in previous research have shown significant differences in the findings of this study. Mean length of T-unit (MLT), mean length of clause (MLC), and coordinate phrases per T-unit (CP/T) did not always yield results that could provide a link between syntactic complexity in L2 writing to L2 proficiency in previous studies. However, MLC has shown significant differences in the current study. Similarly, while coordinate phrases per clause (CP/C) and complex nominals per T-unit (CN/T) have had limited usage in previous research, CP/C has yielded significant results between the two groups. Once again, while these measures do not have enough evidence to reliably measure syntactic complexity in L2 writing, the findings of this study may provide evidence that measuring MLC and CP/T in L2 syntactic complexity research are useful to some extent.

That does not mean that the measures that did not yield significant results are unreliable, however. An important consideration to make here is that the groups used

for this study may not differ very much in terms of English proficiency. Both groups are in their final year of secondary school, so it is possible that they may show less variation than groups of students in the first year when the acquisition of English is still in its early stages.

The current study has some additional limitations that may affect the results. Not much research has gone towards comparing levels of syntactic complexity in L2 written work between CLIL and non-CLIL students, so there are not many results that can be compared to the current findings directly. While previous research of syntactic complexity and L2 writing can be used as a guideline, some differences in the results could be expected. This study also did not consider other individual factors that may advantage CLIL students such as scholastic aptitude, motivational factors, out-of-school contact with English, initial proficiency, and proficiency development. These are only a handful of factors that could influence a student's success at a CLIL programme (Möller, 2017; Verspoor et al., 2015). For example, the students may already be more proficient in English prior to starting the CLIL programme compared to students who choose not to sign up for a CLIL programme at all. Lastly, the dataset used for this study is quite small and is likely that a bigger dataset would have produced more reliable results.

The results are generally in line with previous research that aimed to investigate the differences in L2 proficiency between CLIL and non-CLIL students. As predicted, the CLIL students generally performed slightly better in terms of syntactic complexity compared to non-CLIL students. However, only three out of the eight syntactic complexity measures showed significant differences. While there may be some evidence that CLIL students construct more syntactically complex writing as opposed to non-CLIL students, there are some limitations to this study that may affect results. The dataset for this study was small relative to other studies examining syntactic proficiency in L2 writing, the two groups that were tested may not differ that much in terms of English proficiency, and individual factors such as motivation and pre-CLIL proficiency were not considered.

Conclusion

The aim of this study was to investigate the differences in syntactic complexity in L2 writing of CLIL students compared with non-CLIL students. Two sets of 20 essays written by final-year students of both non-CLIL and CLIL programmes at a secondary school in The Netherlands were analysed, using eight syntactic complexity measures that were used and shown to be reliable in previous research. Similar to the predictions, the essays written by CLIL students generally were more syntactically complex than the essays written by non-CLIL students. CLIL students showed higher mean values in seven out of eight complexity measures (MLS, MLC, MLT, CN/T. CN/C, CP/T, and CP/C), with the exception of clauses per T-unit (C/T). Out of those seven measures, significant results were found in mean length of clause (MLC), coordinate phrases per T-unit (CP/T), and coordinate phrases per clause (CP/C). The results imply that there is some evidence that students in CLIL programmes produce more syntactically complex EFL writing compared to non-CLIL students.

There are some limitations to this study that should be carefully considered, though. While CLIL writing showed significantly higher syntactic complexity than non-CLIL writing in three measures, no significant differences were found in the five other measures. Additionally, the amount of data used in this study was relatively low compared to other research within the field. Furthermore, the groups that were researched were in their final year of secondary school, so it is possible that the two groups did not differ that much in terms of English proficiency. Additionally, individual factors such as student motivation, initial proficiency, and scholastic aptitude that could influence the results other than the distinction of CLIL and non-CLIL were left unexplored.

This study and its results can function as a basis, though, for future research in the field of syntactic complexity in L2 writing of CLIL and non-CLIL students. Future research could benefit from using a larger dataset for more conclusive and reliable evidence. It is most likely beneficial to also analyse individual factors that could influence general proficiency and development, such as the potential bias of learners that are more fit for language-learning to choose to follow a CLIL programme (Möller, 2017). While there are not many developmental studies on L2 success of students in CLIL programmes as opposed to non-CLIL programmes, developmental studies in this setting may help determine the effectivity of CLIL programmes.

Since there is no conclusive evidence on the effectivity of CLIL programmes, more evidence towards proficiency differences (or the lack of) between CLIL and non-CLIL programmes is desired.

Dutch (secondary) school institutions could benefit from the findings of this study. The results provide some insight on the syntactic complexity of written texts by students in CLIL programmes as opposed to texts from students in non-CLIL programmes. This and future studies could help institutions decide on whether to have a CLIL programme based on the results. Since the essays of CLIL students were more syntactically complex, more educational institutes may choose to implement or promote a CLIL programme.

References

Alle tto-scholen in Nederland. (n.d.). Retrieved July 3, 2020, from https://www.nuffic.nl/onderwerpen/alle-tto-scholen-in-nederland/

Admiraal, W., Westhoff, G., & Bot, C. J. L. de. (2006). Evaluation of bilingual secondary education in The Netherlands: Students' language proficiency in English. *Educational Research and Evaluation*, *12*, 75–93.

Bi, P., & Jiang, J. (2020). Syntactic complexity in assessing young adolescent EFL learners' writings: Syntactic elaboration and diversity. *System*, *91*.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. New York: Longman.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5–35.

Biber, D., Gray, B., & Poonpon, K. (2013). Pay Attention to the Phrasal Structures: Going Beyond T-Units-A Response to WeiWei Yang. *TESOL Quarterly*, *47*(1), 192–201.

Biçaku, R. Ç. (2011). CLIL and teacher training. *Procedia - Social and Behavioral Sciences*, *15*, 3821–3825.

Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics (United Kingdom)*, *28*(1), 147–164.

Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, *44*, 51–62.

Daalen, F. van. (2016). *The Effectiveness of Bilingual Education in the Netherlands: Level of English, Accent Preferences and Success.* Leiden University.

de Zarobe, Y. R., & Catalán, R. J. (2009). The receptive vocabulary of EFL learners in two instructional contexts: CLIL versus non-CLIL instruction. In *Content and Language Integrated Learning: Evidence from Research in Europe* (pp. 81–92).

de Zarobe, Y. R. (2010). Written production and CLIL: An empirical study. In *Language use and language learning in CLIL classrooms* (pp. 191–209).

Hamayan, E., & Tucker, G. (1980). Language Input in the Bilingual Classroom and Its Relationship to Second Language Achievement. *TESOL Quarterly*, *14*(4), 453–468.

Housen, A., Folkert, K., & Ineke, V. (2012). Complexity, Accuracy, and Fluency: Definitions, Measurement and Research. In *Gesture* (Vol. 8, pp. 285–301).

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473.

Lambert, C., & Kormos, J. (2014). Complexity, Accuracy, and Fluency in Task-based L2 Research: Toward More Developmentally Based Measures of Second Language Acquisition. *Applied Linguistics*, *35*(5), 607–614.

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, *34*(4), 493–511.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics. *International Journal of Corpus Linguistics*, *15*.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2012). *Automated evaluation of text and discourse with Coh-Metrix*. *Automated Evaluation of Text and Discourse with Coh-Metrix*.

Mitchell, R., Tracy-Ventura, N., McManus, K. (2018). Language Learning during Residence Abroad. In *Anglophone Students Abroad* (pp. 18–51).

Möller, V. (2017). A statistical analysis of learner corpus data , experimental data and individual differences : Monofactorial vs . multifactorial approaches.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044

Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics*, *24*(4), 492-518+558.

Skehan, P. (1996). A Framework for the Implementation of Taskbased Instruction. *Applied Linguistics*, *17*(1), 38–62.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532.

Tavakoli, P., Campbell, C., & Mccormack, J. (2016). Development of Speech Fluency Over a Short Period of Time: Effects of Pedagogic Intervention. *TESOL Quarterly*, *50*(2), 447–471. https://doi.org/10.1002/tesq.244

Verspoor, M., de Bot, K., & Xu, X. (2015). The effects of English bilingual education in the Netherlands. *Journal of Immersion and Content-Based Language Education*, *3*(1), 4–27. https://doi.org/10.1075/jicb.3.1.01ver

Yang, W. (2013). Response to Biber, Gray, and Poonpon (2011). *TESOL Quarterly*, *47*(1), 187–191.